Genetic Heterogeneity in Therapy-Naïve Synchronous Primary Breast Cancers and Their Metastases

Ng et al.

Supplementary Methods

Supplementary References

SUPPLEMENTARY METHODS

Immunohistochemistry

The immunohistochemical profile of the invasive lesions was assessed on 4μm-thick sections, using antibodies against estrogen receptor (ER), progesterone receptor (PR) and HER2 as previously described (1). Positive and negative controls were included in each slide run. The results of ER, PR and HER2 immunohistochemistry were evaluated according to the American Society of Clinical Oncology/ College of American Pathologists guidelines (2, 3).

Histologic grading

Mitotic index was prospectively assessed in the centralized pathologic reviewer as part of the main ESOPE study. In addition, histologic review of tubule formation and nuclear pleomorphism was performed by two pathologists with experience and expertise in breast cancer pathology to determine histologic grade according to the Nottingham grading system (4).

Microdissection and DNA extraction

For all tumor biopsies, 15 eight-μm-thick representative histologic sections of the flash frozen and formalin-fixed paraffin-embedded (FFPE) biopsies of the primary breast cancers and metastases were subjected to microdissection with a needle under a stereomicroscope (Olympus SZ61), to ensure >70% of tumor cell content as previously described (5). Genomic DNA was extracted from each tumor and matched peripheral blood (germline) using the DNeasy Blood and Tissue Kit (Qiagen) and quantified using the Qubit Fluorometer assay (Life Technologies) as previously described (6).

Whole-exome massively parallel sequencing of flash frozen biopsies

Previous sequencing analysis of patient 5 was reported elsewhere (7); independent analyses of the sequencing data and additional genetic analyses of the materials from this patient are reported in the present study. DNA extracted from flash frozen biopsies of the primary breast cancers, metastases and peripheral blood were subjected to whole-exome capture using the SureSelect Human All Exon v4 (Agilent) capture system and to massively parallel sequencing on an Illumina HiSeq 2000 at the

Memorial Sloan Kettering Cancer Center Integrated Genomics Operation (MSKCC IGO) following validated protocols (8-10). An average of 385, 359 and 397 million 75-bp paired-end reads were generated from DNA extracted from primary breast cancers, distant metastases and peripheral blood, respectively, equivalent to median depths of 202x (primary breast cancers, range 189x-229x), 207x (metastasis, range 124x-277x) and 208x (germline, range 58x-267x; **Supplementary Table S2**). Exome sequencing data have been deposited in the Sequence Read Archive under the accession SRP055001.

Whole-exome sequencing data processing was performed as described in Weinreb *et al.* (6). In brief, paired-end reads in FASTQ format were aligned to the reference human genome GRCh37 using Burrows-Wheeler Aligner (v0.7.5a) (11). Local realignment was performed using the Genome Analysis Toolkit (GATK, v2.7.4) (12). PCR duplicates were removed using Picard (v1.92, http://broadinstitute.github.io/picard/). Base quality adjustment was performed using GATK (v2.7.4) (12). Somatic single nucleotide variants (SNVs) were identified using MuTect (v1.0) (13) and somatic small insertions and deletions (indels) were identified using GATK (v2.7.4) (12) and the micro-assembly-based Scalpel (v0.1.1) (14). All indels were manually inspected using the Integrative Genomics Viewer (15). Variants found with >5% global minor allele frequency in dbSNP (Build 137) or that were covered by <10 reads in the tumor or <5 reads in the germline were disregarded. Variants for which the tumor variant allele fraction was <5 times than that of the normal variant allele fraction were disregarded.

Validation of mutations in flash frozen samples and discovery in FFPE samples

Orthogonal validation of mutations found by whole-exome sequencing in the flash frozen samples and mutation discovery in the FFPE samples were performed by either targeted capture massively parallel sequencing using a customized set of baits (EzCap, Nimblegen, Roche) on an Illumina HiSeq 2000 or amplicon sequencing using a custom AmpliSeq panel on an Ion Torrent Personal Genome Machine (PGM) as follows.

Targeted capture sequencing was performed with DNA extracted from the available flash frozen and FFPE tissues and the peripheral blood from cases 1, 4-9, using a custom Nimblegen EzCap bait set targeting all somatic mutations found by exome sequencing in any flash frozen lesion in these seven patients (**Supplementary Table S3**), on an Illumina HiSeq 2000 at the MSKCC IGO following validated protocols (8, 9). Sequencing was performed to a median depth of 493x (range 157x-1,324x) and 312x (range 236x-939x) for the tumors and germline, respectively (**Supplementary Table S2**). Paired-end reads in FASTQ format were aligned to the reference human genome GRCh37 using the Burrows-Wheeler Aligner (v0.7.5a) (11). Local realignment was performed using GATK (v2.7.4) (12). PCR duplicates were removed using Picard (v1.92, http://broadinstitute.github.io/picard/).

Amplicon sequencing was performed with DNA extracted from the available flash frozen and FFPE tissues and peripheral blood from all nine patients using a custom AmpliSeq panel targeting all mutations that could not be validated by targeted capture sequencing for seven patients described above, as well as all somatic mutations identified from exome sequencing of the flash frozen tissues of patients 2 and 3, at the MSKCC IGO (**Supplementary Tables S2 and S3**). Sequencing was performed to a median depth of 670x (range 265x-2,408x) and 1054x (range 690x-2,457x) for the tumors and germline, respectively (**Supplementary Table S2**). Paired-end reads in FASTQ format were aligned to the reference human genome GRCh37 using the Torrent Mapping Alignment Program (v3.4.1, https://github.com/iontorrent/TS/tree/master/Analysis/TMAP). Local realignment was performed using GATK (v3.1.1) (12).

Validation of mutations in the flash frozen samples and discovery of mutations in the FFPE samples were performed using the validation mode of VarScan2 (16) (v2.3.5) and using Scalpel (14) (v0.1.1). Mutations found to be germline variants were excluded from further analysis. Non-germline variants called “Somatic” by VarScan2 or scalpel were considered validated (in the flash frozen tissues) or present (in the FFPE tissues). The median validation rate was 94% (range 47%-99%, with all except one sample above 88%) and the median false negative rate (i.e. not found by whole-exome sequencing but subsequently found to be somatic by targeted capture and/or amplicon sequencing

to be present and somatic) was 1% (range 0%-17%, with all but one sample below 6%, **Supplementary Table S2**). Owing to insufficient DNA, orthogonal validation was not performed for the flash frozen sample of the liver metastasis of patient 5 and for the flash frozen sample of the primary tumor of patient 3. For these two samples, mutations found by whole-exome sequencing that were validated by targeted or amplicon sequencing in at least one related flash frozen or FFPE samples were considered validated. Sufficient DNA could not be obtained from the FFPE biopsy of the primary tumor of patient 5, the FFPE biopsy of the primary tumor of patient 7 and the flash frozen biopsy of the metastasis for patient 8 for the AmpliSeq panel, resulting in a small number of mutations considered "not tested", which are indicated as such in figures. In addition, non-germline variants were considered present by interrogation if they were supported by at least three reads by either targeted capture sequencing or PGM sequencing. Only validated somatic mutations were taken forward for further analysis.

Gene copy number profiling

For the flash frozen biopsies subjected to whole-exome sequencing, FACETS (17) was used to define copy number alterations (CNAs). Specifically, read counts for positions within the target regions with dbSNP entries (build 137) were generated for matched tumor and normal counterpart, and used as input to FACETS, which performs a joint segmentation of the total and allelic copy ratio and infers allele-specific copy number states. DNA extracted from available materials from microdissected FFPE biopsies of the primary breast tumors and metastasis were subjected to copy number profiling analysis using the OncoScan v3 molecular inversion probe array (Affymetrix) following manufacturer's instructions (**Supplementary Table S2**). OncoScan arrays were processed using the OncoScan console. Exported OSCHP files were imported into the Nexus Express for OncoScan software (BioDiscovery, http://www.biodiscovery.com/nexus-express-for-oncoscan/) and processed using the TuScan algorithm within the Nexus Express software.

Segmented Log_2 ratio from FACETS (whole-exome sequencing) and Nexus Express for OncoScan (OncoScan arrays) were used as input for ABSOLUTE (v1.0.6, see also below) (18) to determine

integer copy number and cancer cell fractions (CCFs) of CNAs (19). In brief, ABSOLUTE estimates sample purity, ploidy and absolute copy number by fitting Gaussian mixture models over discrete copy number states over a range of purity and ploidy. After taking into account of the likelihood of the observed karyotype (using data from cytogenetic characterization of human cancer (18)), a set of solutions with maximum likelihood was defined. For the flash frozen biopsies of each patient, minimum and maximum ploidy was set to +/- 1 of the average ploidy estimate by FACETS. The top 3 ABSOLUTE models were retrieved and the pair of solutions with the minimum pairwise distance between their modal copy number estimates was selected. Similarly, for OncoScan arrays, patient-specific minimum and maximum ploidy was set to +/- 1 of the average ploidy estimate by FACETS. For each sample, the top 10 ABSOLUTE solutions were retrieved and the solution with the minimum distance between its modal copy number estimates and the modal copy number estimate of the selected modals for the corresponding flash frozen biopsies (FACETS) was selected. Solutions from ABSOLUTE were manually reviewed as recommended to select a final solution (19). Based on the final solution and its associated mixture models over discrete copy number states, the probability of each CNA being subclonal was estimated (19). CNAs whose subclonal probability is ≥ 0.2 (based on the source code of ABSOLUTE, v1.06) were considered subclonal. Clonal CNAs were assigned CCF 100% and the CCF of subclonal CNAs was computed based on its posterior distribution of CCF values (between 1% and 100%), given the Log_2 ratio and sample purity. As the number of genomic configurations of absolute copy number and CCF increases exponentially with the number of absolute copy number and the tendency of over-fitting of CCFs for focal amplifications, CCF estimates for amplifications (defined below) were disregarded. CCFs for all other types of CNAs were retained.

Gains and losses were defined relative to the average ploidy of all samples from a given patient using the modal copy number for each segment from ABSOLUTE. Segments with modal copy number greater than average ploidy+1 were considered gains, greater than average ploidy+3 amplifications, less than average ploidy-1 losses, modal copy number of 0 homozygous deletions. Copy number states were collapsed based on the median values to cytoband resolution based on the “Chromosome Band (Ideogram)” track from the University of California Santa Cruz Genome Browser

(http://genome.ucsc.edu/index.html). Regions of loss of heterozygosity were defined using FACETS (whole-exome sequencing) and Nexus Express for OncoScan software (OncoScan). For cases where both OncoScan and FACETS results were available (n=6), a substantial to perfect agreement for the CNA profiles was observed (median Cohen's weighted kappa 0.85, range 0.76-0.88; **Supplementary Fig. S1A** and **Supplementary Table S2**) (20).

Identification of likely pathogenic mutations

A combination of MutationTaster (21), CHASM (breast) (22) and FATHMM (23) was used to define the potential functional effect of each missense SNV. Missense SNVs defined as non-deleterious/ passenger by both MutationTaster (21) and CHASM (breast) (22), a combination of mutation function predictors shown to have a high negative predictive value (20), were considered likely passenger alterations. The remaining missense SNVs were defined as likely pathogenic if they were predicted to be "driver" and/ or "cancer" by CHASM (breast classifier) and/ or FATHMM (23), respectively. In-frame indels defined as "neutral" by MutationTaster (21) and PROVEAN (24) were defined as likely passengers. The remaining in-frame indels, as well as frameshift, splice-site and nonsense mutations were considered likely pathogenic if they were targeted by loss of the wild-type allele (i.e. LOH) or affected haploinsufficient genes (25). SNVs, including missense and nonsense SNVs, affecting hotspot residues (17) were also considered likely pathogenic. Mutations were also annotated if they affected genes included in the cancer gene lists described by Kandoth et al. (127 significantly mutated genes) (26), the Cancer Gene Census (27) or Lawrence et al. (Cancer5000-S gene set) (28). Mutations that were neither likely pathogenic nor likely passenger were considered of indeterminate pathogenicity.

Classification of trunk and branch mutations, and mutations enriched in the primary or metastatic lesion

The CCF of each validated mutation in the biopsies of the primary tumor or metastasis was inferred using the number of reads supporting the reference and the alternate alleles obtained from targeted capture or PGM sequencing (or exome sequencing if neither targeted capture or PGM was available)

as secondary input to the copy number analysis using ABSOLUTE (18) (described above). A mutation was classified as clonal if its probability of being clonal was >50% (19) or if the lower bound of the 95% confidence interval of its CCF was >90% (10). Mutations that were considered validated or present by interrogation but do not meet the above criteria were considered subclonal.

A mutation was considered ‘trunk’ if it was found to be clonal in all available biopsies in any given patient. Mutations that were not clonal or were not present in all available biopsies in any given patient were considered ‘branch’. We defined mutations ‘specific to the metastatic lesion’ as those present in at least one biopsy of the metastatic lesion but absent from all biopsies of the primary tumor from the same patient and defined mutations ‘enriched in the metastatic lesion’ as those associated with an increase in CCF by at least 20% in the metastatic lesion compared to the primary tumor, and vice versa for mutations ‘specific to the primary tumor’ and mutations ‘enriched in the primary tumor’.

Analysis of pathways associated with the metastatic process

To identify pathways that may be associated with the metastatic process, we performed pathway analysis on genes affected by likely pathogenic mutations specific to, enriched in the metastatic lesion (see above) and those associated with the loss of the wild-type allele in at least one biopsy of a metastatic lesion and not associated with the loss of the wild-type allele in any biopsy of the matched primary tumor using the Ingenuity Pathway Analysis software (http://www.ingenuity.com) and g:Profiler (29). For the Ingenuity Pathway Analysis, genes were mapped to canonical pathways; P-values ≤ 0.05 were considered significant. For g:Profiler (29), genes were mapped to KEGG and Reactome pathways; P-values ≤ 0.05 after correction using the g:Profiler native method g:SCS were considered significant.

To further enrich for genes associated with the metastatic process, we excluded the genes mutated in >1% of primary invasive breast cancers in the TCGA cohort (30). TCGA invasive breast cancers and their mutations were retrieved from the "Final Full BRCA Sample Summary" and "Mutations - Publicly accessible MAF archives" at https://tcga-data.nci.nih.gov/docs/publications/brca_2012/,

including all non-silent mutations for 463 primary invasive breast cancers, excluding all metastatic lesions. Genes that were targeted by non-silent mutations in >1% of the 463 primary tumors were excluded from the pathway analysis.

Mutational signatures

To define the evolution of mutational signatures, we measured the mutational context of synonymous and non-synonymous trunk SNVs, branch SNVs, SNVs enriched in the primary tumor and SNVs enriched in the metastatic lesion in a given patient, as previously described (10). Mutation sets with fewer than five mutations were excluded. For a given set of SNVs, we classified its dominant mutational signature based on its Pearson's correlation coefficient to the 12 established breast cancer-associated mutational signatures (9, 31). A 1,000-fold bootstrap was performed and assignment to a given mutational signature was based on the number of iterations with the highest Pearson correlation coefficient (10).

Phylogenetic tree construction

A maximum parsimony tree was built for each case using binary presence/ absence matrices based on the repertoire of non-synonymous and synonymous somatic mutations, gene amplifications and homozygous deletions in the biopsies of the primary tumor and the metastatic lesion, as described by Murugaesu *et al.* (32). A starting tree was constructed using the Neighbor-joining method and Hamming distance and optimized using the parsimony ratchet method (33) implemented in the R package Phangorn (34). Trees were rooted at the hypothetical normal where all somatic alterations are absent. Branch lengths were determined according to the ACCTRAN criterion as implemented in the Phangorn package and were drawn to scale.

SUPPLEMENTARY REFERENCES

1. Weigelt B, Ng CK, Shen R, Popova T, Schizas M, Natrajan R, et al. Metaplastic breast carcinomas display genomic and transcriptomic heterogeneity [corrected]. Mod Pathol. 2015;28:340-51.

2. Hammond ME, Hayes DF, Dowsett M, Allred DC, Hagerty KL, Badve S, et al. American Society of Clinical Oncology/College Of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. J Clin Oncol. 2010;28:2784-95.

3. Wolff AC, Hammond ME, Hicks DG, Dowsett M, McShane LM, Allison KH, et al. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. J Clin Oncol. 2013;31:3997-4013.

4. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology. 1991;19:403-10.

5. Natrajan R, Wilkerson PM, Marchio C, Piscuoglio S, Ng CK, Wai P, et al. Characterization of the genomic features and expressed fusion genes in micropapillary carcinomas of the breast. J Pathol. 2014;232:553-65.

6. Weinreb I, Piscuoglio S, Martelotto LG, Waggott D, Ng CK, Perez-Ordonez B, et al. Hotspot activating PRKD1 somatic mutations in polymorphous low-grade adenocarcinomas of the salivary glands. Nat Genet. 2014;46:1166-9.

7. Bidard FC, Ng CK, Cottu P, Piscuoglio S, Escalup L, Sakr RA, et al. Response to dual HER2 blockade in a patient with HER3-mutant metastatic breast cancer. Ann Oncol. 2015;26:1704-9.

8. Kohsaka S, Shukla N, Ameur N, Ito T, Ng CK, Wang L, et al. A recurrent neomorphic mutation in MYOD1 defines a clinically aggressive subset of embryonal rhabdomyosarcoma associated with PI3K-AKT pathway mutations. Nat Genet. 2014;46:595-600.

9. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500:415-21.

10. Schultheis AM, Ng CK, De Filippo MR, Piscuoglio S, Macedo GS, Gatius S, et al. Massively Parallel Sequencing-Based Clonality Analysis of Synchronous Endometrioid Endometrial and Ovarian Carcinomas. J Natl Cancer Inst. 2016;108:djv427.

11. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754-60.

12. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297-303.

13. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013;31:213-9.

14. Narzisi G, O'Rawe JA, Iossifov I, Fang H, Lee YH, Wang Z, et al. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. Nat Methods. 2014;11:1033-6.

15. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178-92.

16. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568-76.

17. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat Biotechnol. 2016;34:155-63.

18. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30:413-21.

19. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell. 2013;152:714-26.

20. Martelotto LG, Ng C, De Filippo MR, Zhang Y, Piscuoglio S, Lim R, et al. Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. Genome Biol. 2014;15:484.

21. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7:575-6.

22. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. 2009;69:6660-7.

23. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34:57-65.

24. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012;7:e46688.

25. Dang VT, Kassahn KS, Marcos AE, Ragan MA. Identification of human haploinsufficient genes and their genomic proximity to segmental duplications. Eur J Hum Genet. 2008;16:1350-7.

26. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502:333-9.

27. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer. 2004;4:177-83.

28. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014;505:495-501.

29. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. 2016;44:W83-9.

30. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61-70.

31. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. 2016;534:47-54.

32. Murugaesu N, Wilson GA, Birkbak NJ, Watkins TB, McGranahan N, Kumar S, et al. Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy. Cancer Discov. 2015;5:821-31.

33. Nixon KC. The Parsimony Ratchet, a new method for rapid parsimony analysis. Cladistics-the International Journal of the Willi Hennig Society. 1999;15:407-14.

34. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27:592-3.